AFRL-RI-RS-TR-2017-224

CREATING ROBUST RELATION EXTRACT AND ANOMALY DETECT VIA PROBABILISTIC LOGIC-BASED REASONING AND LEARNING

UNIVERSITY OF WISCONSIN

NOVEMBER 2017

FINAL TECHNICAL REPORT

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE

■ **AIR FORCE MATERIEL COMMAND** ■ **UNITED STATES AIR FORCE** ■ **ROME, NY 13441**

NOTICE AND SIGNATURE PAGE

AFRL-RI-RS-TR-2017-224 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE CHIEF ENGINEER:

/ S /	**/ S /**
JAMES M. NAGY	MICHAEL J. WESSING
Work Unit Manager	Deputy Chief, Information Intelligence
	Systems and Analysis Division
	Information Directorate

REPORT DOCUMENTATION PAGE	*Form Approved* **OMB No. 0704-0188**

1. REPORT DATE *(DD-MM-YYYY)*
NOVEMBER 2017

2. REPORT TYPE
FINAL TECHNICAL REPORT

3. DATES COVERED *(From - To)*
OCT 2012 – AUG 2017

4. TITLE AND SUBTITLE

CREATING ROBUST RELATION EXTRACT AND ANOMALY DETECT VIA PROBABILISTIC LOGIC-BASED REASONING AND LEARNING

5a. CONTRACT NUMBER
FA8750-13-2-0039

5b. GRANT NUMBER
N/A

5c. PROGRAM ELEMENT NUMBER
62303E

6. AUTHOR(S)

Jude Shavlik, Chris Re, Sriraam Natarajan

5d. PROJECT NUMBER
DEFT

5e. TASK NUMBER
12

5f. WORK UNIT NUMBER
16

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Wisconsin
1300 University Avenue
Madison, WI 53706

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Research Laboratory/RIED
525 Brooks Road
Rome NY 13441-4505

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/RI

11. SPONSOR/MONITOR'S REPORT NUMBER
AFRL-RI-RS-TR-2017-224

12. DISTRIBUTION AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited. This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09

13. SUPPLEMENTARY NOTES

14. ABSTRACT

We consider a three-pronged approach to deep exploration and filtering of text. The first is development of a set of scalable state-of-the-art learning algorithms that are capable of learning generalized probabilistic logic rules from noisy, incomplete data. The second is a data management system that is widely accepted as the state-of-the art for knowledge base construction (KBC) and is highly scalable. The final direction is the design and adaptation of the scalable management and learning algorithms for the tasks of deep knowledge understanding such as knowledge-based population and anomaly detection.

In this report, they organize and present their accomplishments (the approaches and their intuitive, theoretical and empirical ramifications) from the DEFT cooperative agreement into 3 main focus areas or research thrusts. Each of them is motivated and introduced separately in their respective sections.

15. SUBJECT TERMS
Robust Relational extractors, anomaly detectors, Probabilistic logic-based reasoning, reasoning and learning, machine learning, relationship extraction, implicit information understanding, Natural language understanding, NLU, belief and modality, statistical inference

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** **JAMES M. NAGY**
a. REPORT U	**b. ABSTRACT** U	**c. THIS PAGE** U	UU	42	**19b. TELEPHONE NUMBER (*Include area code)*** **N/A**

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

We acknowledge and thank all the members of the StaRAI research laboratory and all the collaborators who have been instrumental, directly or indirectly, in our progress and accomplishments.

1 SUMMARY

The key research accomplishments out of the DEFT cooperative agreement are three fold:

1. **Development of an efficient learning algorithm for relational probabilistic models:** Historically, Artificial Intelligence has used either the logical approach (to address structured problems) or the statistical approach (to handle uncertainty). Recent years have witnessed a tremendous development of techniques to handle large-scale, structured and uncertain domains. Statistical Relational Learning (SRL) considers the problem of learning in the presence of rich, multi-relational, semi-structured data. While these models are highly attractive due to their compactness and comprehensibility, the task of learning in these models is computationally intensive. As part of the DEFT cooperative agreement, we developed an efficient learning algorithm for learning directed, undirected and bi-directed SRL models in the presence of hidden data/missing values, a problem that has rarely been addressed before in the community that traditionally makes a closed-world assumption to handle missing data.

2. **Development of the state-of-the-art data management system for automatic knowledge base construction:** One key challenge in building a high-quality knowledge base construction (KBC) system is that developers must often deal with data that are both diverse in type and large in size. Further complicating the scenario is that these data need to be manipulated by both relational operations and state-of-the-art machine-learning techniques. This research thrust focuses on supporting this complex process of building KBC systems. DeepDive is a data management system that we built to study this problem; its ultimate goal is to allow scientists to build a KBC system by declaratively specifying domain knowledge without worrying about any algorithmic, performance, or scalability issues.

3. Design, implementation, and evaluation of algorithms for several deep language understanding tasks such as knowledge-based population and anomaly detection: The basic scientific research pursued under this cooperative agreement were aimed at solving deep Natural Language Processing (NLP) problems. In particular, we focused on knowledge base population towards the later years and initially focused on anomaly detection from text data.

In the subsequent sections, we present the accomplishments in detail and provide empirical support for our algorithms. Finally, we will conclude by listing all the papers/manuscripts and book published as a result of funding from this cooperative agreement.

2 INTRODUCTION

We consider a three-pronged approach to deep exploration and filtering of text. The first is development of a set of scalable state-of-the-art learning algorithms that are capable of learning generalized probabilistic logic rules from noisy, incomplete data. The second is a data management system that is widely accepted as the state-of-the art for KBC and is highly scalable. The final direction is the design and adaptation of the scalable management and learning algorithms for the tasks of deep knowledge understanding such as knowledge-based population and anomaly detection.

In this report, we organize and present our accomplishments (the approaches and their intuitive, theoretical and empirical ramifications) from, the Deep Exploration and Filtering of Text (DEFT) cooperative agreement into 3 main focus areas or research thrusts. Each of them is motivated and introduced separately in their respective sections.

3 METHODS, ASSUMPTIONS AND PROCEDURES

3.1 Effective Statistical Relational Learning

3.1.1 Motivation

Statistical Relational Learning (SRL) studies the combination of relational learning (e.g. inductive logic programming) and statistical machine learning. By combining the power of logic and probability, such approaches can perform robust and accurate reasoning and learning about complex relational data. The advantage of these formulations is that they can succinctly represent probabilistic dependencies among the attributes of different related objects, leading to a compact representation of learned models. Most of these methods essentially use first-order logic to capture domain knowledge and soften the rules using probabilities or weights.

While statistical relational models are indeed highly attractive due to their compactness and comprehensibility, learning them is typically much more demanding than learning propositional ones. Most of these methods essentially use first-order logic to capture domain knowledge and soften the rules using probabilities or weights. At least this was the goal of the models that were developed initially in this area. This is due to the fact that Inductive Logic Programming as a field was evolving then. This field concerns mainly with the problem of learning first-order rules from data. So early systems simply used a logic learner underneath to learn the rules and simply employed probabilistic learning techniques such as maximum likelihood estimation (for complete data) and Expectation-Maximization (EM) (for incomplete data) to learn the parameters (i.e., weights or probabilities) for these rules. This is due to the fact learning structure of a SRL model requires learning the parameters repeatedly in the inner loop which in turn can sometimes require probabilistic inference in its inner loop. This problem is exaggerated for SRL models since the predicates can allow for arbitrary combinations of variables or constants as arguments[1]. Consequently many early SRL methods relied of human experts providing the structure and only learned the parameters of models.

Parameter learning reduced the amount of effort needed from the expert and potentially improved the accuracy of the model (by relying on data to correct mistakes made by experts), but also increased the computational time. For some domains, the structure of the model may be non-trivial, not known or insufficient. As a result, both the structure and parameters of the model need to be learned from the data.

Although *structure learning* reduces the expert's effort, it can be computationally intensive due to the large space of possible structures while including parameter learning as a sub-task. Figure 1 shows this trade-off between the computation cost and expert's effort. This project focuses on reducing the learning time while improving the accuracy of structure learning in SRL models

[1] For instance, when learning to predict if someone (say x) is popular, it is possible to use the predicate *Friends* in several ways. Some possible ways are *Friends(x,y), Friends(y,x), Friends(x,"Erdos")* and *Friends("Erdos",x).* Of course, the constant "Erdos" can be replaced with all possible constants in the data base

So, is scaling up learning of SRL models insurmountable? No, as we will argue here. Typical SRL approaches seek to learn a single best SRL model. However, if learning a single best and easy-to-interpret model is difficult, we should maybe stop acting as if our goal is to do so, and instead embrace (model) complexity: we turn the problem of learning a single SRL model into a series of relational regression problems learned in a stage-wise manner using Friedman's functional gradient boosting.

This is a sensible idea since finding many rough rules of thumb of how to change a model can be a lot easier than finding a single model.

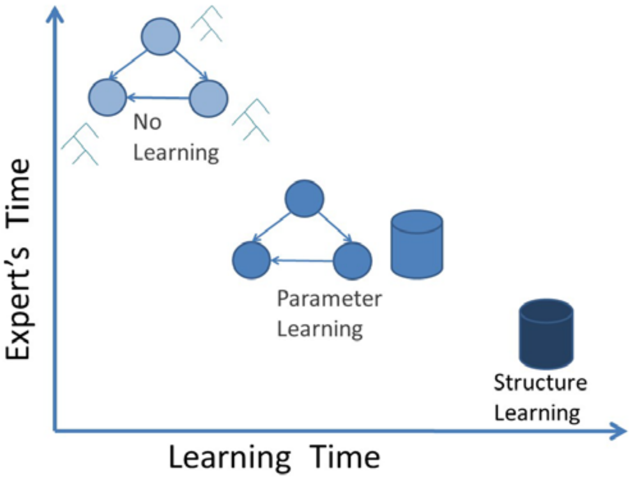


Figure 1: Trade-offs on learning problems in SRL models. expert's time vs. learning time

3.1.2 Research Accomplishments

An important advancement in our research thrust is triggered by the insight that finding many rough rules of thumb of how to change our probabilistic relational models locally can be a lot easier than finding a single, highly accurate local model. Consider for example relational dependency networks. Instead of learning the conditional probability distribution associated with each predicate using relational tree learning in single shot manner, one can represent it as a weighted sum of regression models grown in a stage-wise optimization using gradient *boosting*.

The key idea in this line of work is to represent each conditional distribution as a set of relational regression trees (RRT), Figure 2 and Figure 3 illustrate an example of RRT. Figure 2 shows a relational functional gradient boosting. Relational Functional gradient boosting is similar to the standard gradient-boosting where trees are induced in stage-wise manner. At every iteration, the gradients are computed as the difference between observed and predicted probabilities of each example and a new regression tree is fitted to these examples. Figure 3 is an example of a RRT. The goal is to predict if a person A is advisedBy B (where A and B are logical variables) given the properties and relations of people at a university. At each node a (set of) predicate(s) is evaluated. The leaves denote regression values that are exponentiated and normalized to obtain the probabilities. The second branch from the left states that if B is a professor, A is not a professor,

A has more than 1 publication, and has more than 1 common publication with B, then the regression value is 0.05. These regression values are then exponentiated and normalized.

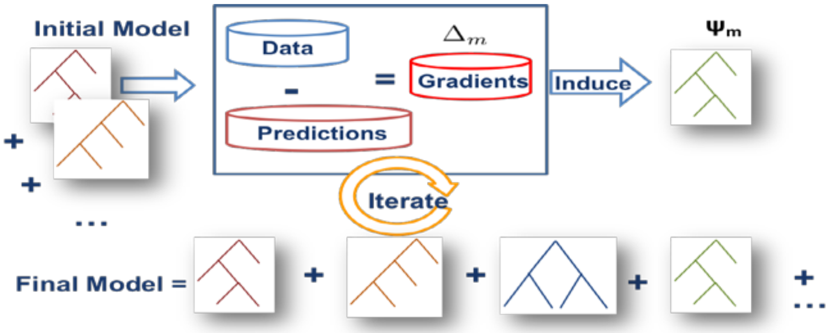


Figure 2: Relational Functional Gradient Boosting

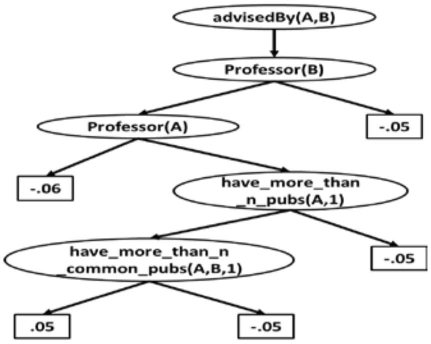

advisedBy(A,B)
Professor(B)
Professor(A)
-.05
-.06
have_more_than_n_pubs(A,1)
have_more_than_n_common_pubs(A,B,1)
-.05
.05
-.05

Figure 3: Example of a Relational Regression Tree

To learn this set of RRTs for each conditional distribution, the learning approach was adapted based on the gradient boosting technique and was called relational functional-gradient boosting (RFGB). Assume that the training examples are of the form $(\mathbf{x}_i, y_i)$ for $i = 1, \ldots, N$ and $y_i \in \{1, \ldots, K\}$. Let us use $\mathbf{x}$ to denote the set of non-target predicates (features) and y_i to denote the current target predicate.

Then the goal is to fit a model $P(y|\mathbf{x}) \propto e^{\psi(y,\mathbf{x})}$.

The key idea is to compute the gradient (weight) for each example separately and fit a regression tree over all the weighted examples. This set of local gradients will approximate the global gradient.

The functional gradient of each example $(\mathbf{x}_i, y_i)$ w.r.t likelihood ($\psi(y_i = 1; \mathbf{x}_i)$) is

$$\frac{\partial \log P(y_i; \mathbf{x}_i)}{\partial \psi(y_i = 1; \mathbf{x}_i)} = I(y_i = 1; \mathbf{x}_i) - P(y_i = 1; \mathbf{x}_i) \qquad (1)$$

where I is the indicator function that is 1 if $y_i = 1$ and 0 otherwise. The expression is simply the adjustment required to match the predicted probability with the true label of the example. If the example is positive and the predicted probability is less than 1, this gradient is positive indicating that the predicted probability should move towards 1. Conversely, if the example is negative and the predicted probability is greater than 0, the gradient is negative, driving the value the other way. They use RRTs to fit the gradient function for every training example.

Each RRT can be viewed as defining several new feature combinations, each corresponding to one of the paths from the root to a leaf. The resulting potential functions from all these different RRTs still have the form of a linear combination of features but the features can be quite complex. This idea is illustrated in Figure 3.

The benefits of a boosted learning approach are manifold.

- First, being a nonparametric approach the number of parameters grows with the number of training episodes. In turn, interactions among random variables are introduced only as needed, so that the potentially large search space is not explicitly considered.
- Second, such an algorithm is fast and straightforward to implement. Existing off-the-shelf regression learners can be used to deal with propositional, continuous, and relational domains in a unified way.
- Third, it learns the structure and parameters simultaneously, which is an attractive feature as learning probabilistic relational models is computationally quite expensive.

For an implementation as well as further material on how to use it, we refer to http://pages.cs.wisc.edu/~tushar/rdnboost/index.html.

While originally developed for relational dependency networks, this work was later extended to learning MLNs by simply optimizing the pseudo-likelihood.

Handling hidden data. Inspired by the success of structural EM on propositional graphical models and the success of boosting in learning SRL models, we developed an EM algorithm for functional-gradient boosting (FGB). We derive and present the update equations of the E and M-steps of the algorithm. One of the key features of our algorithm is that we consider the set of distributions in the models to be a product of potentials and this allows us to learn different models such as Markov Logic Networks (MLNs) and relational dependency networks (RDNs). After deriving the EM algorithm, we adopt the standard approach of approximating the full likelihood by the maximum a posteriori (MAP) states (i.e., hard EM). We show that this MAP approximation to the likelihood makes learning computationally tractable with minimal change in quality. As far as we are aware, this is the first work on combining EM with FGB for relational domains.

Let us denote the observed data using **X** and the hidden data using **Y**. Also, let us use 1 and 0 to represent true and false respectively. Given a training set with missing data, the goal is to maximize the log likelihood of the observed groundings. We average the likelihood function over all possible world states of the missing data (joint assignment over all hidden groundings) to compute the marginal probabilities of the observed groundings as shown below.

$$\pounds(\psi) \equiv \log P(\mathbf{X} = \mathbf{x} \mid \psi) \qquad \text{Likelihood } (\pounds) \text{ of observed data } \mathbf{x}$$

$$= \log \Sigma\; P(\mathbf{x};\mathbf{y} \mid \psi) \qquad \text{Marginalize over hidden data instantiations}$$

$$\mathrm{y}\text{€}\mathrm{Y}$$

$$= \log \sum_{\mathrm{y}\text{€}\mathrm{Y}} \{P(\mathbf{y}|\mathbf{x};\psi^{!}) \quad \frac{P(\mathbf{x};\mathbf{y}|\psi)}{P(\mathbf{y}|\mathbf{x};\psi t)}\} \qquad \text{Multiply and divide by} P(\mathbf{y}\,\mathbf{x};\psi^{!}) \qquad (2)$$

Assume $\psi^{!}$ is our current estimate of the best model based on the log-likelihood function. We will derive gradient steps to find ψ that has a higher log-likelihood than $\psi^{!}$. We then set the ψ obtained via these gradient steps as the new $\psi^{!}$ and iteratively find a better ψ. To make the iterative procedure clearer, we use ψ_t to represent the $\psi^{!}$ obtained after t iterations of the gradient steps.

We present the high-level overview of our RFGB-EM (Relational Functional Gradient Boosting - EM) approach in Figure 4. Similar to EM approaches that we explained earlier, we sample the states for the hidden groundings based on our current model in the E-step and use the sampled states to update our model in the M-step. ψ_t represents the model in the t^{th} iteration. The initial model, ψ_0 can be as simple as a uniform probability for all examples or could be a model specified by an expert. We sample certain number of assignments of the hidden groundings (denoted as $|W|$) using the current model ψ_t. Based on these samples, we create regression examples which are then used to learn T relational regression trees. The learned regression trees are added to the current model and the process is repeated. We refer to our paper [16,25] for the detailed algorithm and the mathematical details.

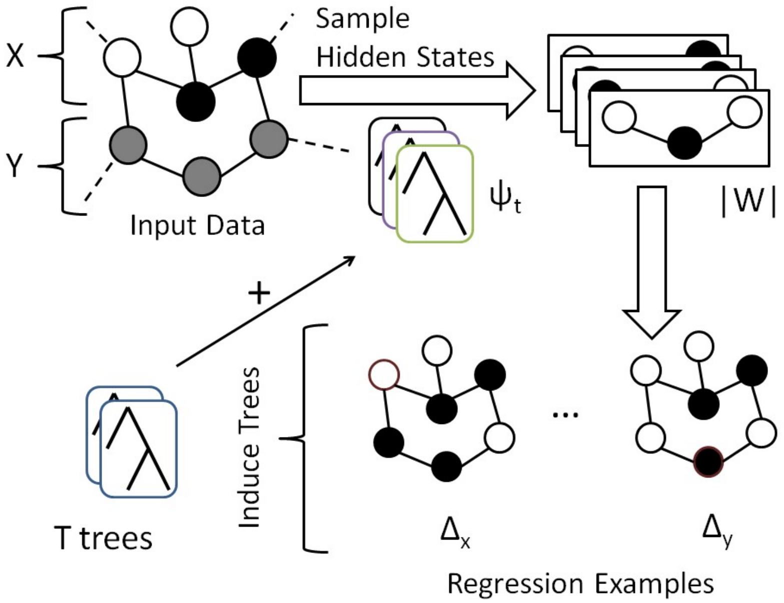


Figure 4: RFGB-EM flowchart

In figure 4, the Shaded nodes indicate variables with unknown assignments, while the white (or black) nodes are assigned true (or false) values. The input data has observed (indicated by X) and hidden (indicated by Y) groundings. We sample $|W|$ assignments of the hidden groundings using the current model ψ_t. We create regression examples based on these samples, which are used to learn T relational regression trees. The learned trees are added to the current model and the process is repeated.

Handling imbalanced data. We extended the framework to handle is misclassification costs on imbalanced class distributions where over-sampling or under-sampling do not work due to overfitting. Imbalanced class distributions have remained a significant bottleneck for relational algorithms due to several relations being false.

For instance, consider `Friends(x,y)`. If `x` and `y` can take 1000 values each, this relation can have 1M different grounded facts, while only a fraction of them will be true in the world. We defined a cost function,

$c(x_i) = \alpha I(y_i = 1 \wedge y = 0) + \beta I(y_i = 0 \wedge y = 1),$

where $I(y_i = 1 \wedge y = 0)$ is 1 for false negatives

and $I(y_i = 0 \wedge y = 1)$ is 1 for false positives.

Intuitively, $c(x_i) = \alpha$ when a positive example is misclassified,

while $c(x_i) = \beta$ when a negative example is misclassified.

Now, the gradients were derived and the learning was performed as earlier. Our results showed that the proposed method was robust to handling severe class imbalance in a principled manner [42].

Learning from advice. Our framework is also capable of treating human as more than a mere labeler. To this effect, we developed a framework capable of handling preferences as advice. We aim to learn a model that has higher probabilities for examples that satisfy advice as compared to ones that do not. Consider s to be the set of training examples for which the advice is applicable,

i.e., $\mathbf{s} = \{s_i \mid B, F$ f-- $\mathrm{label}(s_i)\}$, where

B is the background knowledge and F is the advice constraint.

The learned model should have a higher probability for the label that satisfies the advice than the probability of the label that violates the advice, i.e.,

$$\forall s_i \in \mathbf{s}, P(l + (s_i)) \geq P(l - (s_i)).$$

We propose to incorporate this advice into RFGB by modifying the objective function to include a cost,

$$MLL(\mathbf{x},\mathbf{y}) = \sum \log \frac{e^{\psi(x_i; y)})}{\mathrm{cost}(x_i,\psi) + \sum e^{\psi(x_i; y)}} \qquad (3)$$

More details of this work can be seen in our AAAI paper [33]. This work was later applied to the task of medical relation extraction from text [32].

Learning from one-class data. To solve the problem of learning with examples from one class, techniques such as Support Vector Machines (SVMs), nearest-neighbors and clustering have been proposed. These methods rely on a distance measure between the examples being provided. For propositional data, standard distance measures are available and such measures can also be learned. But data in real world domains tends to be more complicated and structured, which can naturally be represented using first-order logic where standard distance measures are not applicable. Hence, we developed an approach to perform one- class classification with a novel distance measure for relational data. This approach can be used to perform anomaly analysis as well as information extraction on rich structured data as seen in Natural Language Processing tasks.

Relational examples can be viewed as examples with infinite dimensions (since there are infinite first- order rules where each rule can be used to create a boolean feature). The key idea of our approach is to select relevant dimensions from these infinite dimensions that keep examples of the same class close to each other while spreading out the unlabeled examples (similar to the idea in Principal Component Analysis approach for dimensionality reduction). Since our approach is a feature selection method, we can also use it for one-class classification in very high-dimensional data. Rather than using first-order logic rules to propositionalize the relational examples and learn a distance measure, we directly learn a relational distance function using trees. We use dissimilarity between the paths taken by examples in the relational trees to compute the distance between the examples. We learn multiple trees where each tree contributes to the distance between examples. We refer to our AAAI paper [17] for more details.

Efficient SRL using relational databases. Statistical Relational Learning approaches learn accurate models from noisy structured data but combining probabilities to model noise and first order logic to model the relational structure. The limitations of most SRL models stem from logical operations and aggregations they need to perform over multiple iterations, at times over large amounts of data, which affects their efficiency. This issue can potentially be alleviated by exploiting the power of relational databases. Recent database centric systems such as tuffy have made considerable progress in that direction, but are limited to only parameter learning. We develop and formulate an approach that fully integrates relational databases with the Inductive Logic Programming engine that performs structure learning in SRL. Substantial speed-up is achieved via representing facts as im-memory database tables instead of text files and translating logical and aggregation operations into equivalent SQL queries. Additional efficiency gains result from precomputing and storing count data and caches for reoccurring costly join operations. Our evaluations prove the effectiveness of our approach paired with the state-of-the-art SRL framework RFGB. WE refer to our ILP 2016 best paper [21] for further details.

Scaling via approximation in Relational Models. Inference in statistical relational models is computationally expensive since full instantiation of first order theories is exponentially large. Lifted inference tries to alleviate this problem by performing inference over groups of instances that are equivalent/symmetric (induce the same belief on the query). But, even lifted inference techniques suffer from scaling issues for large datasets since it requires counting over number of instances in such equivalence groups. In learning, counting is an essential operation as a part of the recurring inference in structure learning and as a part of likelihood estimation in parameter learning. In a first order logic representation of the relational structures, counting is essentially over the satisfied instances of a FOL clause given the data. Estimating satisfiability via brute force is combinatorial search problem and is #P -complete. However, exact counting is not essential for large datasets since minor deviations have negligible effect on final likelihood or belief estimates. We developed an approach to approximate the counts of satisfied instances by compiling the data into a graph database representation and using a message passing technique to estimate expected values of counts from summary statistics of the constructed graph. Our approach FACT has demonstrated substantial efficiency gains on large datasets without compromising the performance in terms of learning or inference in statistical relational models. We refer to our SDM paper [6] for more details.

Learning via Domain Adaptation/Transfer. Sophisticated machine learning models allow for effective and accurate learning from large amounts of data. With limited dataset sizes, however, accurate modeling is difficult. The advent of transfer learning techniques alleviate this limitation by using knowledge (model) gained from a source task for better and more effective learning in a target task with limited amount of data. Most techniques focus on transferring across related tasks in the same domain and do not work well for tasks in unrelated domains. Cross-domain transfer necessitates relational and structured representation. Recent transfer learning techniques for relational domains induce higher order generalized knowledge (such as 2nd-order logic) from the source domain and use that knowledge to learn better models in a target domain. While a reasonable approach, higher order logic is complicated to reason with and learning is less efficient. We proposed and developed Language-biased Transfer Learning (LTL) that performs domain-constrained walks on the source model (eg.: a FOL theory) and induces an initial model for the target domain aligned with the walks from the source domain, which is then refined via the small amount of data available for the target domain. The constrained walks are formally termed as Mode-Matching Trees (M 2T s) which implicitly capture the knowledge due to the domain language itself, including how variables are shared across different relations and how the source theory would be instantiated. A single path on the source domain tree may then induce several ones in the target tree which can later be refined, allowing better convergence to the most accurate theory in the target domain. Intuitively, it can be viewed as tracing the shape of the source domain ontology and ensuring that the target model adheres to that shape as mush as possible, making the SRL structure search both effective and efficient. Empirical evaluations prove the effectiveness of our method in transferring across seemingly unrelated domains (eg: Sports and Finance, NELL ontologies). We refer to our ICDM paper [19] for further details and discussion.

Relational Deep models. Connectionist approaches, including Restricted Boltzman Machines (RBM), have become increasingly popular for learning probability distributions due to their expressive power which has let to their success in varied tasks such as collaborative filtering, motion capture, density modeling of images and speech and so on. Being a connectionist framework, RBMs allow for 'stacking', leading to deep models. However, it suffers from the limitations of interpretability and the necessity of data being represented as finite dimensional flat feature vectors (i.i.d.). We formulated and developed Relational RBMs that enable RBMs to operate on noisy, structured data.

The key idea behind our proposed formalism is to construct relational features via random walks (inspired by the Path Ranking Algorithm) on the relational schema (object/entity classes and relations between them) and using the support for such features (counts/existentials) as inputs to a discriminative RBM (since we are interesting in classification tasks). We show how our approach relates to Statistical Relational Learning approaches. The relational feature construction via random walks is akin to structure learning of relational models (eg. MLN clauses) and the discriminative RBM acts as the parameter learning layer for the relational structure. This, in turn, provides some degree of interpretability to RBMs. We demonstrate how our RRBMs exhibit better or equal performance to state-of-the-art SRL approaches (RDN-Boost/MLN-Boost) on several relational datasets. We refer to our ILP paper [15] for further details.

3.2 DeepDive Developments

Our development of DeepDive has proceeded as a series of iterations of *application*, *technique*, and *abstraction*. Applications from domain scientists provide the motivation and prioritize a series of core computer science challenges; these challenges become the focus of a series of studies of relevant techniques, and these techniques are then made available to the user via a declarative abstraction that in turn facilitates more applications.

3.2.1 Application and Quality: PaleoDeepDive

Paleontology is based on the description of fossils. Unfortunately, most fossil data are buried in journal articles published over the past four centuries. To make these data usable by paleontologists, nearly two decades ago, a team of more than 300 scientists spent nearly 10 continuous person years manually reading 40K publications to compile one of the largest paleontological knowledge bases, PaleoBioDB, which has been the basis of more than 200 scientific publications. This stunning amount of human effort and the huge scientific impact of PaleoBioDB raised the question: *Can we apply our prototype system built for DARPA to automate this process but still achieve human levels of quality?* This motivated our collaboration with geoscientist Shanan Peters to build PaleoDeepDive [34]. The goal was to take as input journal articles and construct knowledge bases in the same schema as PaleoBioDB. We conducted a series of studies on the quality of this process.

Rule-based vs. Inference-based Systems. State-of-the-art KBC researches are usually conducted along one of two different lines—those based on hard rules and those based on statistical inference. We conducted studies [29, 30] to investigate the impact of statistical inference on the quality of KBC. We found that, with similar engineering efforts and enough training examples, across a range of five different applications, the inference-based approach often achieves higher quality. This is not surprising, because for the inference-based approach, the user only needs to specify what features are potentially relevant and let the inference algorithm decide their quality (weights); however, for a rule-based system, the developer must deal with both. This decoupling of relevancy and quality allows users to add many relevant but possibly noisy sources to boost the quality of a KBC system quickly.

Directly Supervised vs. Distantly Supervised Systems. As our study about inference-based KBC systems reveals, for an inference system to be able to make decisions about the quality of features, it is important to have a large set of training examples. Thus, We investigated different ways of creating such examples. As for textbook supervised machine-learning systems, many such training examples can be manually labeled by experts, which results in high-quality training examples, but this process is usually expensive and time-consuming. Instead, We studied [47] two less expensive but potentially lower quality, noisy ways of generating training examples, namely distant supervision and crowd-sourcing. Distant supervision allows the user to write rules to generate training examples from an unstructured corpus by linking it to an existingKB, while crowd-sourcing allows the user to delegate the labelling procedure to non-professional workers. This study found that distant supervision can achieve better quality given a large enough unstructured corpus, and the key insight is that the noise produced by the supervision procedure can be mitigated by a large number of training examples with statistical learning.

Single-modality vs. Multi-modality Systems. The above studies formed the basis for building the first version of PaleoDeepDive, which extracts information from sources such as text and tables separately. However, this first prototype had a less than satisfying quality—To understand a table inside a document, we often need to consult information in other tables or text in the rest of the document. This motivated a follow-up study [11] in which we designed a joint inference scheme to conduct inference concurrently across text and tables. This study shows the significant impact of such a joint inference on the quality of PaleoDeepDive; however, it introduced statistical inference tasks that often require terabytes of data, which pose challenges for the scalability and performance of the system.

Discussion With this engine and the series of study on quality, for all fossil-related relations in Paleo-BioDB, PaleoDeepDive achieves comparable, and sometimes better, quality than human volunteers [34] and forms the foundation of another dozen of applications [2, 22, 34, 45, 52] that follow similar approaches.

3.2.2 Performance and Scalability: Scalable Statistical Inference and Learning

The array of studies shown above demonstrate the necessity for a scalable and efficient engine that can conduct joint statistical inference and learning. Compared with a traditional relational workload, one key difference of a statistical workload, as we found in many applications [49], is the decoupling of *hardware efficiency* and *statistical efficiency*. That is, an implementation that takes full advantage of the hardware might converge slowly from a statistical perspective, and thus lead to suboptimal end-to-end efficiency compared with a more balanced implementation. Worse, there is only a limited body of theory to guide the choice between these two angles. Thus, our research takes a *systems approach* for both scalability and efficiency.

Scalable Statistical Inference with Gibbs Sampling. As statistical inference is often #P-hard in general, one workhorse approximate algorithm is Gibbs sampling. We conducted a study [48] on running Gibbs sampling over data that does not fit in the main memory. our approach is to revisit three classic database techniques that are used in storage managers: materialization, page-oriented layout, and buffer-replacement policy. After studying the tradeoffs associated with the different choices of these techniques, We developed a prototype system that achieves an increase in speed of up to two orders of magnitude over traditional baseline approaches. This study also revealed the potential of adapting classic database techniques to this new workload after a systematic revisit.

Performant Main-memory Statistical Analytics. We then conducted a study on how to speed up the execution of inference inside the main-memory buffer. The need for this type of research is also relevant to an industrial trend; today, even small organizations have access to machines with large main memories. Not surprisingly, there has been a flurry of activity to support main-memory analytics in both industry and research. However, each of these systems often picks one design point in a larger tradeoff space. Thus, our research [49] seeks to define and study this tradeoff space, focusing on commodity multi-socket, multi-CPU, non-uniform memory access (NUMA) machines. This tradeoff space contains axes such as data access methods, data replication, and model replication. From this study, We found that today's research and industrial systems often underutilize commodity modern hardware for analytics, sometimes by two orders of magnitude. This study results in a system called DimmWitted, the workhorse inference engine inside DeepDive that supports all its execution of statistical inference algorithms.

3.2.3 Abstraction and Usability: Iterative Feature Engineering

The above studies, along with our other work in systems [13, 31] and *collaborative* work in theory [7, 8, 40, 50], forms the cornerstone of an efficient DeepDive engine. One remaining challenge is determining what abstraction and interface DeepDive should provide to the user. One observation we made by observing scientists' use patterns is that building high-quality KBC systems is not a one-shot process; instead, it is an iterative process in which the user keeps trying different combinations of features and executes statistical inference on different but similar tasks. Therefore, We focused on studying an iterative abstraction for two popular workloads: feature selection and feature engineering.

Iterative Feature Selection. Feature selection is the process of selecting a set of features that will be used to build a statistical model—a process that is widely regarded as the most critical step of statistical analytics. We found that [46] feature selection is an interactive human-in-the-loop process. For this reason, we designed a declarative language to specify a feature selection workload. This declarative language and the iterative model mean that feature selection workloads are rife with reuse opportunities. Thus, we studied how to materialize portions of this computation, using not only classical database optimizations but also methods that have not previously been used in databases, including structural decomposition methods and warmstart. This study found that traditional database-style approaches that ignore these new opportunities are more than two orders of magnitude slower than an optimal plan in this new tradeoff space across multiple execution backends, and that a simple cost-based optimizer can often automatically select a near-optimal execution plan for feature selection.

Iterative Feature Engineering. The workload of developing a KBC system is often more complicated than just feature selection. In most applications built with DeepDive, we have seen a spectrum of changes that can be made during the development—quality requirements change, new data sources arrive, and new concepts are needed in the application. This finding motivates our desire to develop techniques for making the entire pipeline incremental in the face of changes, both to the data and to the DeepDive program. The first component of this study is a language similar to Datalog that allows the user to specify feature extraction, distant supervision, and statistical inference and learning using a unified, relational, and declarative language. The main technical challenge is to incrementally maintain statistical inference and learning given a changed factor graph. We developed two methods for incremental inference based respectively on sampling and variational techniques, and we further study the tradeoff space of these methods in order to develop a simple rule-based optimizer [38]. These techniques speed up KBC inference tasks by up to two orders of magnitude with negligible impact on quality. This work forms DeepDives current language and interaction model, which most of our applications build on.

3.3 Deep NLP Tasks

We now describe our deep NLP tasks that we have addressed in the past three years.

3.3.1 Applications and Challenges

Knowledge base construction (KBC) is the process of populating a knowledge base with facts extracted from unstructured data sources such as text, tabular data expressed in text and in structured forms, and even maps and figures. In *sample-based science* [35], one typically assembles a large number of facts (typically from the literature) to understand macroscopic questions, e.g., about the amount of carbon in the Earth's atmosphere throughout time, the rate of extinction of species, or all the drugs that interact with a particular gene. To answer such questions, a key step is to construct a high-quality knowledge base, and some sciences have undertaken decade-long sample collection efforts, e.g., PaleoDB.org and PharmaGKB.org.

In parallel, KBC has attracted interest from industry [10,51] and academia [3–5,9,14,18,24,28,36,37,41, 43]. To understand the common patterns in KBC systems, we are actively collaborating with scientists from a diverse set of domains, including geology [44], paleontology [35], pharmacology for drug repurposing, and others. We first describe one KBC application we built, called PaleoDeepDive, then present a brief description of other applications built with similar purposes.

PaleoDB and PaleoDeepDive Paleontology is based on the description and biological classification of fossils, an enterprise that has been recorded in and an untold number of scientific publications over the past four centuries. One central task for paleontology is to construct a knowledge base about fossils from scientific publications, and an existing knowledge base compiled by human volunteers has greatly expanded the intellectual reach of paleontology and led to many fundamental new insights into macroevolutionary processes and the nature of biotic responses to global environmental change. However, the current process of using human volunteers is usually expensive and time-consuming. For example, PaleoDB, one of the largest such knowledge bases, took more than 300 professional paleontologists and 11 human years to build over the last two decades, resulting in `PaleoDB.org`. To get a sense of the impact of this database on this field, at the time of writing, this dataset has contributed to 205 publications, of which 17 have appeared in *Nature* or *Science*.

This provided an ideal test bed for our KBC research. In particular, we constructed a prototype called PaleoDeepDive [35] that takes in PDF documents. This prototype attacks challenges in optical character recognition, natural language processing, information extraction,

Beyond Paleontology The success of PaleoDeepDive motivates a series of other KBC applications in a diverse set of domains including both natural and social sciences. Although these applications focus on very different types of KBs, they are usually built in a way similar to PaleoDeepDive. This similarity across applications motivate our study of building DeepDive as a unified framework to support these diverse applications.

Human-Trafficking MEMEX is a DARPA program that explores how next generation search and extraction systems can help with real-world use cases. The initial application is the fight against human trafficking. In this application, the input is a portion of the publicly-indexed and ”dark” web in which human traffickers are likely to (surreptitiously) post supply and demand information about illegal labor, sex workers, and more. DeepDive processes such documents to extract evidential data such as names, addresses, phone numbers, job types, job requirements, information about rates of service, etc. Some of these data items are difficult for trained human annotators to accurately extract and have never been previously available, but DeepDive- based systems have high accuracy (Precision and Recall in the 90s, which may surpass that of non-experts). Together with provenance information, such structured, evidential data are then passed on to both other collaborators on the MEMEX program as well as law enforcement for analysis and consumption in operational applications. MEMEX has been featured extensively in the media and is supporting actual investigations. For example, every human trafficking investigation pursued by the Human Trafficking Response Unit in New York City now involves MEMEX, for which DeepDive is the main extracted data provider. In addition, future use cases such as applications in the war on terror are under active consideration.

Medical Genetics The body of literature in life sciences has been growing at an accelerating speed, to the extent that it has been unrealistic for scientists to perform research solely based on reading and/or keyword search. Numerous manually-curated structured knowledge bases are likewise unable to keep pace with exponential increases in the number of publications available online. For example, OMIM is an authoritative database of human genes and mendelian genetic disorders which dates back to the 1960s, and so far contains about 6,000 hereditary diseases or phenotypes, growing at a rate of roughly 50 records / month for many years. Conversely, almost 10,000 publications were deposited into PubMed Central per month last year. In collaboration with Prof. Gill Bejerano at Stanford, we are developing DeepDive applications to create knowledge bases in the field of medical genetics. Specifically, we use DeepDive to extract mentions of genes, gene variants, and phenotypes from the literature, and statistically infer their relationships, presently being applied to clinical genetic diagnostics & reproductive counseling.

Pharmacogenomics Understanding the interactions of chemicals in the body is key for drug discovery. However, the majority of this data resides in the biomedical literature and cannot be easily accessed. The Pharmacogenomics Knowledgebase is a high quality database that aims to annotate the relationships between drugs, genes, diseases, genetic variation, and pathways in the literature. With the exponential growth of the literature, manual curation requires prioritization of specific drugs or genes in order to stay up to date with current research. In collaboration with Emily Mallory and Prof. Russ Altman [23] at Stanford, we are developing DeepDive applications in the field of pharmacogenomics. Specifically, we use DeepDive to extract relations between genes, diseases, and drugs in order to predict novel pharmacological relationships.

TAC-KBP TAC-KBP is a NIST-sponsored research competition where the task is to extract common properties of people and organizations (e.g., age, birthplace, spouses, and shareholders) from a 1.3 million newswire and web documents – this task is also termed Slot Filling. In the 2014 evaluation, 31 US and international teams participated in the competition, including a solution based on DeepDive from Stanford [1]. The DeepDive based solution achieved the highest precision, recall, and F1 among all submissions.

3.3.2 Dependency Networks for KBP

Our Approach: We present our system for KBP slot filling based on probabilistic logic formalisms and present the different components of the system. Specifically, we employ Relational Dependency Networks, a formalism that has been successfully used for joint learning and inference from stochastic, noisy, relational data. We consider our RDN system against the current state-of-the-art for KBP to demonstrate the effective- ness of our probabilistic relational framework. Additionally, we show how RDNs can effectively incorporate many popular approaches in relation extraction such as joint learning, weak supervision, `word2vec` features, and human advice, among others.

We provide a comprehensive comparison of various settings such as joint learning vs learning of individual relations, use of weak supervision vs gold standard labels, using expert advice vs only learning from data, etc. These questions are extremely interesting from a general machine learning perspective, but also critical to the NLP community. As we show empirically, the key contributions of this paper are as follows:

- Our RDN framework is competitive, and often superior, to state-of-the-art systems for KBP slot filling.
- RDNs successfully incorporate various types of features, including advice, joint learning, and `word2vec` features.
- Ours is the first KBP system to leverage *knowledge-based weak supervision* – a logic-based framework that we have previously shown to be complementary and often superior to distant supervision.

Some of the results such as human advice being useful in many relations and joint learning being beneficial in the cases where the relations are correlated among themselves are on the expected lines. However, some surprising observations include the fact that weak supervision and `word2vec` features are not as useful as expected, although further investigation is warranted.

Given a training corpus of raw text documents, our learning algorithm (Figure 5) first converts these documents into a set of facts (i.e., features) that are encoded in first order logic (FOL). Raw text is processed using the Stanford CoreNLP Toolkit[2] to extract parts-of-speech, word lemmas, etc. as well as generate parse trees, dependency graphs and named-entity recognition information. The full set of extracted features are then converted into features in prolog (i.e., FOL) format and are given as input to the system.

[2] http://stanfordnlp.github.io/CoreNLP

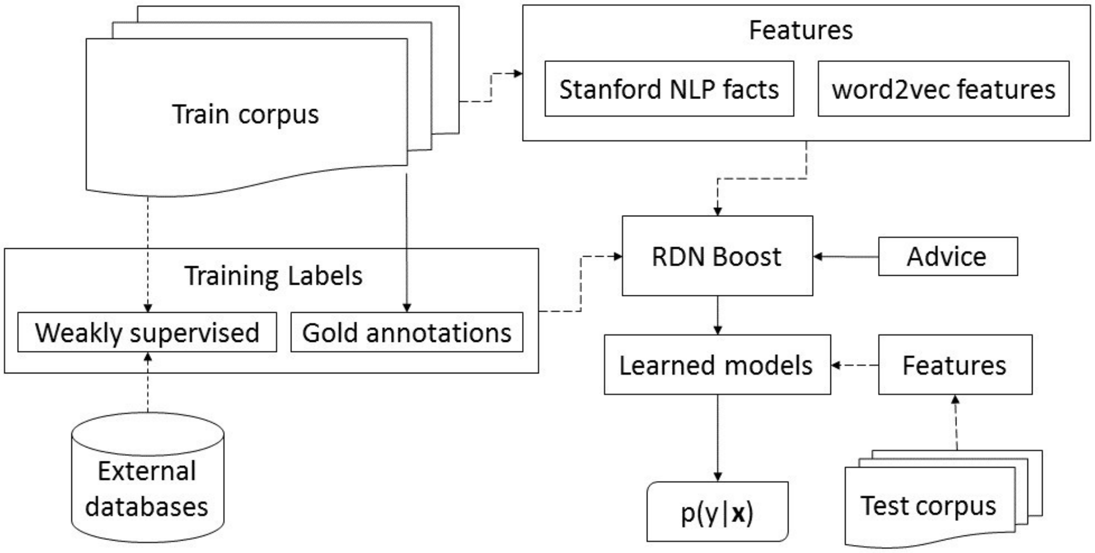


Figure 5: Pipeline: Full RDN relation extraction pipeline

In addition to the structured features from the output of Stanford toolkit, we also use deeper features based on `word2vec` as input to our learning system. Standard NLP features tend to treat words as individual objects, ignoring links between words that occur with similar meanings or, importantly, similar contexts (e.g., city-country pairs such as *Paris* – *France* and *Rome* – *Italy* occur in similar contexts). `word2vec` provide a continuous-space vector embedding of words that, in practice, capture many of these relationships. We use word vectors from Stanford[3] and Google[4].

We generated features from word vectors by finding words with high similarity in the embedded space. That is, we used word vectors by considering relations of the following form:

isCosSimilar(*wordA, relationB, threshold*), if a word has a high cosine similarity to any keyword (e.g., “father”) for a particular relation (e.g., *parent*). We set the modes such that the first argument to *isCosSimilar* is a ‘-’ (i.e., existing) variable, while the second and third arguments are constants. We provide lists of candidate constants (on average a few dozen for each target relation) for the second argument using our knowledge of each target concept. For example, we include words such as “father” and “mother” (inspired by the *parent* relation) or “devout”,“convert”, and “follow” (*religion* relation).

One difficulty with the KBP task is that very few documents come labeled as *gold standard labels*, and further annotation is prohibitively expensive beyond a few hundred documents. This is problematic for discriminative learning algorithms, like the RDN learning algorithm, which excel when given a large supervised training corpus. To overcome this obstacle, we employ *weak supervision* – the use of external knowledge (e.g., a database) to heuristically label examples. We employ our novel knowledge-based weak supervision approach, as opposed to the more traditional distant supervision.

[3] http://nlp.standford.du/projects/glove/
[4] https://code.google.com/p/word2vec/

An alternative approach, knowledge-based weak supervision is based on previous work [27,39] with the following insight: labels are typically created by "domain experts" who annotate the labels carefully, and who typically employ some inherent rules in their mind to create examples. For example, when identifying family relationship, we may have an *inductive bias* towards believing two persons in a sentence with the same last name are related, or that the words "son" or "daughter" are strong indicators of a parent relation. We call this *world knowledge* as it describes the domain (or the world) of the target relation.

Table 1: Rules for KB Weak Supervision.

Weight	MLN Clause
1.0	entityType(a, "PER"), entityType(b, "NUM"), nextWord(a, c), word(c, ","), nextWord(c, b) → age(a, b)
0.6	entityType(a, "PER"), entityType(b, "NUM"), prevLemma(b, "age") → age(a, b)
0.8	entityType(a, "PER"), entityType(b, "PER"), nextLemma(a, "mother") → parents(a, b)
0.8	entityType(a, "PER"), entityType(b, "PER"), nextLemma(a, "father") → parents(a, b)

For the KBP task, some rules that we used are shown in Table 1. Table 1 is a sample of knowledge-based rules for weak supervision. The first value defines a weight or confidence in the accuracy of the rule. The target relation appears at the end of each clause. "PER", "ORG", "NUM" represent entities that are persons, organizations, and numbers, respectively. For example, the first rule identifies any number following a person's name and separated by a comma is likely to be the person's age (e.g., "Sharon, 42"). The third and fourth rule provide examples of rules that utilize more textual features; these rules state the appearance of the lemma "mother" or "father" between two persons is indicative of a parent relationship. Previous results show this approach produces more examples with less overhead than distant supervision and can be employed where relevant database are not available.

To this effect, we encode the domain expert's knowledge in the form of first-order logic rules with accompanying weights to indicate the expert's confidence. We use the probabilistic logic formalism *Markov Logic Networks* to perform inference on unlabeled text (e.g., the TAC KBP corpus). Potential entity pairs from the corpus are queried to the MLN, yielding (weakly-supervised) positive examples. We choose MLNs as they permit domain experts to easily write rules while providing a probabilistic framework that can handle noise, uncertainty, and preferences while simultaneously ranking positive examples. We use the Tuffy system to perform inference as it is robust and scales well to millions of documents[5].

[5] As the structure and weights are predefined by the expert, learning is not needed for our MLN

3.3.3 Extracting Medical Relations from Text

While there is a plethora of research on detecting Adverse Drug Events (ADEs) from clinical data, there are not many methods that can validate the output of these algorithms except for manually scanning through the ADEs. In this case, the burden is on the expert to evaluate these extracted ADEs by knowing all the ones published in the literature. We explore the use of published medical abstracts to serve as ground truth for evaluation and present a method for effectively extracting the ADEs from published abstracts. To this effect, we adapt and apply our advice-based algorithms presented earlier that use a human expert (say a physician) as more than a "mere labeler", i.e., the human expert in our system is not restricted to merely specify which of the drug event pairs are true ADEs. Instead, the human expert would "teach" the system much like a human student by specifying patterns that he/she would look for in the papers. These patterns are employed as advice by the learning system that seamlessly integrates this advice with training examples to learn a robust classifier.

More precisely, given a set of ADE pairs (drug-event pairs), we build upon an NLP pipeline to rank the ADE pairs based on the proof found in the literature. Our system first searches for PubMed abstracts that are relevant to the current set of ADE pairs. For each ADE pair, these abstracts are then parsed through a standard NLP parser (we use Stanford NLP parser) and the linguistic features such as parse trees, dependency graphs, word lemmas and n-grams etc. are generated. These features are then used as input to our RFGB algorithm for learning to detect ADEs from text. While powerful, standard learning will not suffice for the challenging task of extracting ADEs as we will show empirically. The key reason is that we do not have sufficient number of training examples to learn a robust classifier. Also, the number of linguistic features can be exponential in the number of examples and hence learning a classifier in this hugely imbalanced space can possibly yield sub-optimal results. To alleviate this imbalance and guide the learner to a robust prediction model, we explore the use of human guidance as advice to the algorithm. This advice could be in terms of specific patterns in text. For instance it is natural to say something like, "if the phrase *no evidence* is present between the drug and event in the sentence then it is more likely that the given ADE is not a true ADE". The learning algorithm can then identify the appropriate set of features (from the dependency graph) and make the ADE pair more likely to be a negative example.

The drug and event pairs come from Observational Medical Outcomes Partnership [6] 2010 ground truth, a manually curated database. To facilitate evaluation and comparison of methods and databases, OMOP established a common data model so that disparate databases could be represented uniformly. This included definitions for ten ADE-associated health outcomes of interest (HOIs) and drug exposure eras for ten widely-used classes of drugs.

Since this OMOP data includes very few positive examples (10 to be precise), we investigated other positive examples found in the literature to increase the training set. Our final dataset that we built contains 39 positive and 1482 negative examples (i.e., 39 x 38, the cross-product of all drug-effect pairs and obtained the ones that are not true ADE). The abstracts that we collected for the drug and event pairs contained 5198 sentences. Note that some drug and event pairs were not mentioned in any abstracts. In all experiments, we performed 4-fold cross validation. We compare both area under the curve for ROC and PR curves. We discuss the empirical results in later sections.

[6] http://omop.org/

3.3.4 Anomaly Detection from Text

We consider a supervised approach to identifying anomalies in text. Our definition of anomaly also follows the standard definition of "deviating from normal (expected) situations". We are interested in document classification, i.e., identifying documents that deviate from the normal. We hypothesize that the definition of anomaly in the context of textual data depends on the domain of interest. For instance, when reading sports articles, an example anomaly is when a low-ranked team playing away from home defeats a top-ranked team. This "upset" can be identified based on the knowledge of the teams, their relative rankings etc. and not necessarily on the lexical features. Similarly, in the recent unfortunate incident of the missing flight, the size of the aircraft and the zone in which it was flying are crucial to identifying it as an anomaly. When tagging anomalies in a question forum, the context of a question (for example, requesting solutions to a homework problem) is crucial. The common theme across all these situations is that specific information about the domain is more important than the lexical features. Such domain knowledge can be naturally provided in first-order logic (FOL) as features or advice rules. Consequently, it is easier to learn using richer representations such as Statistical Relational Learning (SRL). We employ a recently successful learning algorithm called relational functional gradient-boosting (RFGB) [26] for learning to predict the anomalies and show that it outperforms standard approaches.

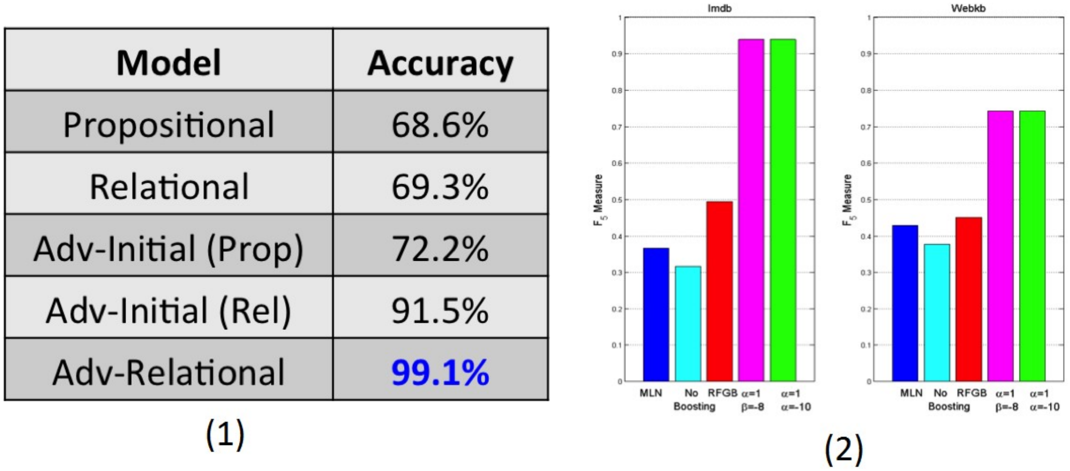
We made a few key contributions in this work [20]: (1) we show that using domain knowledge can substantially improve the detection of anomalies in text and (2) we also show that simply looking at syntactic features can greatly reduce the predictive performance of automated anomaly detection (3) finally, we evaluate using two domains - a literature domain, inspired by the work of Guthrie [12], where the goal is to identify text that does not belong to a particular author, and a flight domain where the goal is to read about flight incidents and identify the relatively "unexpected" incidents.

4 RESULTS AND DISCUSSION

We discuss our results and findings based the focus areas of research and our methods and approaches that we enumerate above.

4.1 Effective SRL via advice

Figure 6 presents some sample results of learning from advice in a segmentation task and on standard SRL domains with high class imbalance. The results clearly demonstrate that our advice-based approach learns effectively better models even on not just relational data sets, but propositional ones as well. In segment(1), Adv-initial are the methods that use advice as prior knowledge and learn from them using propositional (prop) and relational (rel) classifiers. For segment (2), In standard PLM domains using misclassification costs. The last 2 sets of columns are different α and β values.



Model	Accuracy
Propositional	68.6%
Relational	69.3%
Adv-Initial (Prop)	72.2%
Adv-Initial (Rel)	91.5%
Adv-Relational	99.1%

Figure 6: Sample results of learning from advice
(1) in a segmentation task and (2)In standard PLM domains

Figure 7 illustrates the success of our advice-based SRL framework on a real-world task of predicting Adverse Drug Events (ADEs) from text, specifically medical abstracts. We defer the detailed discussion about the ADE results to a later stage. Except for alchemy, the systems evaluated were developed by project members.

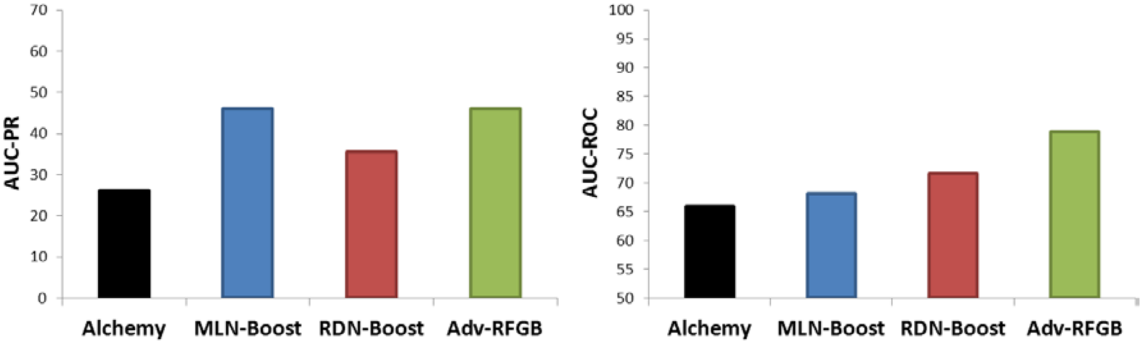


Figure 7: Experimental results for predicting ADEs

4.2 DeepDive and Deep NLP

PaleoDB & PaleoDeepDive Some statistics about our prototype PaleoDeepDive process are shown in Figure 8. As part of the validation of this system, we performed a double-blind experiment to assess the quality of the system versus the PaleoDB. We found that the KBC system built on DeepDive has achieved comparable—and sometimes better—quality than a knowledge base built by human volunteers over the last decade [35] and leads to similar scientific insights on topics such as biodiversity. This quality is achieved by iteratively integrating diverse sources of data- often quality scales with the amount of information we enter into the system. Figure 8 illustrates the accuracy of the results in PaleoDeepDive.

Quality of PaleoDeepDive

Relation	PDD # Ext.	PDD Accuracy	Human Accuracy
Taxon-Taxon	27 M	97%	92%
Taxon-Fm.	4 M	96%	84%
Fm.-Time	3 M	92%	89%
Fm.-Location	5 M	94%	90%

Quality of Other Applications

Applications	F1 Score
Human Trafficking	72%
TAC-KBP	34%
Medical Genetics	53%
Pharmacogenomics	57%

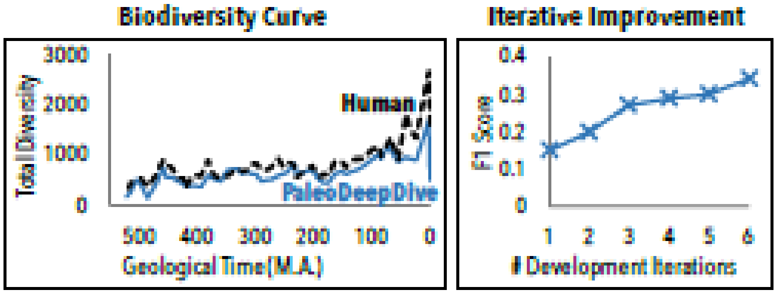


Figure 8: Quality of KBC system built with DeepDive.

Dependency Networks for KBP For the experimental evaluation of the Knowledge Base Population (KBP) task via our dependency network approach we considered 14 specific relations from two categories, *person* and *organization* from the TAC KBP competition. The relations considered are listed in the left column of Table 2. In Table 2, the Columns indicate the number of training examples utilized – both human annotated (Gold) and weakly supervised (WS), when available – from TAC KBP 2014 and number of test examples from TAC KBP 2015. 10 relations describe person entities (*per*) while the last 4 describe organizations (*org*). We utilize documents from KBP 2014 for training while utilizing documents from the 2015 corpus for testing.

Table 2: The relations considered from TAC KBP

Relation	Gold	WS	Test
$per : age$	89	750	44
$per : alternateName$	28	x	18
$per : children$	89	x	23
$per : origin$	96	750	48
$per : otherFamily$	72	750	10
$per : parents$	71	750	30
$per : religion$	70	750	11
$per : siblings$	77	750	31
$per : spouse$	66	750	28
$per : title$	158	x	39
$org : cityHQ$	69	x	10
$org : countryHQ$	69	21	29
$org : dateFounded$	70	750	17
$org : foundedBy$	62	750	32

All results presented are obtained from 5 different runs of the train and test sets to provide more robust estimates of accuracy. We consider three standard metrics – area under the ROC curve, F-1 score and the recall at a certain precision. We chose the precision as 0.66 since the fraction of positive examples to negatives is 1:2 (we sub-sampled the negative examples for the different training sets). It must be mentioned that not all relations had the same number of hand-annotated (gold standard) examples because the 781 documents that we annotated had different number of instances for these relations. The train/test gold-standard sizes are provided in the table, including weakly supervised examples. Negative examples are created by randomly selecting paired entities in the same sentence.

Our system performs comparably, and often better than the state-of-the-art RelationFactory system (Table 3 located in the conclusions section of this report). In particular, our method outperforms RelationFactory in AUC ROC across all relations. Recall provides a more mixed picture with both approaches showing some improvements – RDN outperforms in 6 relations while RelationFactory does so in 8. The values in bold indicate superior performance against the alternative approach. Note that in the instances where RDN provides superior recall, it does so with dramatic improvements (RF often returns 0 positives in these relations). F1 also shows RDN's superior performance, outperforming RF in most relations. Thus, our RDN framework performs comparably, if not better, across all metrics against the state-of-the-art.

Table 3: RelationFactory vs RDN

Relation	AUC ROC		Recall		F1	
	RF	RDN	RF	RDN	RF	RDN
age	0.64	**0.93**	0.28	**0.74**	0.44	**0.67**
alternateName	0.50	**0.77**	0.00	**0.16**	0	**0.10**
children	0.54	**0.76**	0.09	**0.14**	0.17	**0.28**
origin	0.50	**0.89**	0.00	**0.86**	0	**0.64**
otherFamily	0.56	**0.90**	**0.11**	0.06	**0.24**	0.22
parents	0.29	**0.74**	**0.33**	0.15	**0.50**	0.31
religion	0.50	**0.81**	0	**0.56**	0	**0.60**
siblings	0.13	**0.81**	**0.17**	0.00	0.29	0.29
spouse	0.57	**0.85**	**0.13**	0.04	0.23	**0.37**
title	0.67	**0.90**	**0.67**	0.07	**0.80**	0.54
cityHQ	0.38	**0.74**	**0.38**	0.28	**0.55**	0.41
countryHQ	0.57	**0.77**	0.14	**0.62**	0.25	**0.58**
dateFounded	0.67	**0.86**	**0.33**	0.05	**0.50**	0.46
foundedBy	0.20	**0.84**	**0.37**	0.25	0.54	**0.55**

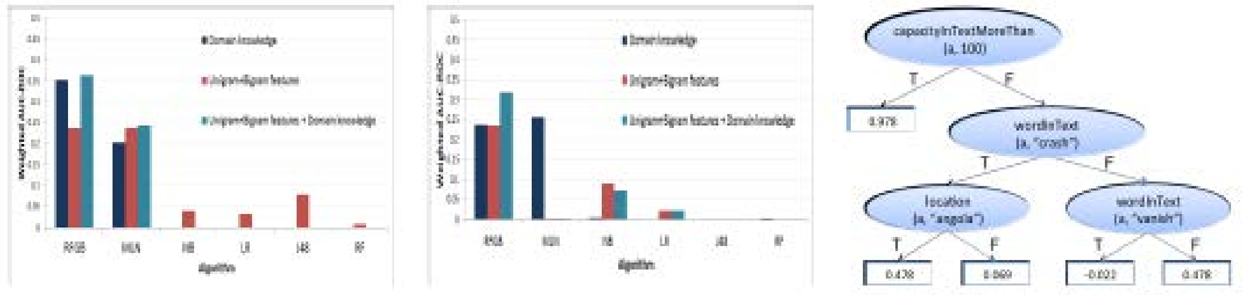
Extracting medical relations from text The results for extracting medical relations from text (specifically ADEs) are presented in figure 7 above. The first three graphs present the results of using only data and employing standard relational learning methods. As can be seen, our proposed method that also employs "human advice" outperforms the three baselines that do not incorporate advice - (*RDN-Boost*, *MLN-Boost*, and *Alchemy*). This highlights the high value that the expert knowledge can have when learning with few training examples. Also, our proposed method is effectively learning with a high degree of accuracy to predict from the text abstracts. It is also clear that the advice is effectively incorporated when compared to merely using the data for learning and inference.

We investigated the differences between our predictions and the OMOP ground truth to understand whether our method was truly effective. One key example where our method predicted an ADE pair to be positive, but OMOP labeled it as a negative ADE pair was: **Bisphosphonate** causes **Acute Renal Failure**. Our method predicted it as an ADE with a high (98.5%) probability. We attempted to validate our prediction and were able to find evidence in the literature to support our prediction. As an example, PubMed article (PMID 11887832) contains the sentence:

Bisphosphonates have several important toxicities: **acute renal failure**, worsening renal function, reduced bone mineralization, and osteomalacia.

This suggests that our method (1) is able to find some evidence to support its prediction and (2) is capable of incorporating novel medical findings. We refer to our paper for more details on our system [32].

Anomaly Detection To evaluate our approach for anomaly detection from text we chose our first domain as the identification of *anomalous flight incidents* from text. We created a dataset consisting of 45 news articles reporting different flight incidents. Anomalies in this domain refer to unexpected flight incidents such as missing or crashed passenger aircrafts with more than 100 passengers in non-war zones. Of the 45 articles, we identified 18 articles as anomalous. Inspired by Guthrie's work [12] on anomaly detection in a literature domain, we created a second dataset with anomalous literary excerpts. We selected excerpts with 20 – 30 sentences from Sir Arthur Conan Doyle's books to create the set of normal documents. To create anomalous documents, we introduced a sentence, at random, from Jane Austen's books in some excerpts from Doyle. We ran three-fold cross validation and present the results in Figure 9. As can be seen, in both the domains, the relational methods (RFGB and MLN) in all the three settings outperform the propositional methods significantly when measuring the weighted AUC-ROC. This indicates that relational methods were more effective in identifying the anomalous text. This result is in line with several previously published works of employing these relational learning methods.



(a) (b) (c)

Figure 9: Results of the experiments in: (a) Flight Domain and (b) Literature Domain. (c) A learned tree in the flight domain

5 CONCLUSION

Our research thrusts including developing efficient learning algorithms that can operate on complex structured data. To this effect, we pursued research in two directions - that of improving probabilistic data bases and developing an effective learning algorithm for probabilistic relational models. We focused our application to deep NLP tasks while evaluating and iterating over the learning algorithms.

In our first direction, we presented the first-of-its-kind learning algorithm for SRL models based on functional gradient boosting. The key idea is to consider learning as a series of regression models in a stage-wise manner. The advantage of this approach is that we can learn the structure of the models and their parameters simultaneously. We showed how to adapt this algorithm for learning multiple relational models

- RDNs, MLNs, relational policies via imitation learning and even transfer learning. We also extended the algorithm to learn faithfully in the presence of missing data without making assumptions about the data. Finally, we demonstrated across several tasks including few different health care prediction problems, natural language processing tasks, transfer learning in relational domains etc. The variety of applications clearly show demonstrate the broad applicability and usefulness of the boosting approach to learning in relational problems.

For our second direction, we introduced DeepDive, a data management system to make knowledge base construction easier. We described a declarative language and an iterative debugging protocol that a user can use to build KBC systems, a set of high-quality KBC systems built with DeepDive, and techniques that we developed to make DeepDive scalable and efficient. These techniques optimize both batch execution and iterative execution, both of which are necessary, according to our experience observing users using DeepDive. With DeepDive, we hope to free domain experts from manually constructing knowledge bases, which is often tedious, expensive, and time consuming.

Our final direction was demonstrating how the two related yet different directions were applied to deep NLP tasks with the focus on KBP systems. Our initial results are promising and we shall continue to focus on tedious tasks that users are currently performing manually for their research and seek clean ways to automate them.

6 REFERENCES

[1] Gabor Angeli et al. Stanford's 2014 slot filling systems. *TAC KBP*, 2014.

[2] Gabor Angeli, Sonal Gupta, Melvin Johnson Premkumar, Christopher D Manning, Christopher Ré, Julie Tibshirani, Jean Y Wu, Sen Wu, and Ce Zhang. Stanford's distantly supervised slot filling systems for KBP 2014. In *Text Analysis Conference Proceedings*, 2015.

[3] Michele Banko et al. Open information extraction from the Web. In *IJCAI*, 2007.

[4] Justin Betteridge, Andrew Carlson, Sue Ann Hong, Estevam R Hruschka Jr, Edith LM Law, Tom M Mitchell, and Sophie H Wang. Toward never ending language learning. In *AAAI Spring Symposium*, 2009.

[5] Andrew Carlson et al. Toward an architecture for never-ending language learning. In *AAAI*, 2010.

[6] Mayukh Das, Yuqing Wu, Tushar Khot, Kristian Kersting, and Sriraam Natarajan. Scaling lifted probabilistic inference and learning via graph databases. In *SDM*, 2016.

[7] Christopher De Sa, Ce Zhang, Kunle Olukotun, and Christopher Ré. Taming the wild: A unified analysis of Hogwild!-style algorithms. *NIPS*, 2015.

[8] Christopher De Sa, Ce Zhang, Kunle Olukotun, and Christopher Ré. Rapidly mixing Gibbs sampling for a class of factor graphs using hierarchy width. *NIPS*, 2015. **Spotlight**.

[9] Oren Etzioni et al. Web-scale information extraction in KnowItAll: (preliminary results). In *WWW*, 2004.

[10] David Ferrucci et al. Building Watson: An overview of the DeepQA project. *AI Magazine*, 2010.

[11] Vidhya Govindaraju, Ce Zhang, and Christopher Ré. Understanding tables in context using standard NLP toolkits. In *ACL*, 2013.

[12] D. Guthrie. *Unsupervised detection of Anomalous Text*. PhD thesis, University of Sheffield, 2008.

[13] Stefan Hadjis, Firas Abuzaid, Ce Zhang, and Christopher Ré. Caffe con Troll: Shallow ideas to speed up deep learning. In *DanaC*, 2015.

[14] Gjergji Kasneci et al. The YAGO-NAGA approach to knowledge discovery. *SIGMOD Rec.*, 2009.

[15] Navdeep Kaur, Gautam Kunapuli, Tushar Khot, Kristian Kersting, William Cohen, and Sriraam Natarajan. Relational restricted boltzmann machines: A probabilistic logic learning approach. In *ILP*, 2017.

[16] Tushar Khot, Sriraam Natarajan, Kristian Kersting, and Jude W. Shavlik. Gradient-based boosting for statistical relational learning: the markov logic network and missing data cases. *Machine Learning*, 100(1):75–100, 2015.

[17] Tushar Khot, Sriraam Natarajan, and Jude W Shavlik. Relational one-class classification: A non-parametric approach. In *AAAI*, pages 2453–2459, 2014.

[18] Rajasekar Krishnamurthy et al. SystemT: A system for declarative information extraction. *SIGMOD Rec.*, 2009.

[19] Raksha Kumaraswamy, Phillip Odom, Kristian Kersting, David Leake, and Sriraam Natarajan. Transfer learning via relational type matching. *2015 IEEE International Conference on Data Mining*, 2015.

[20] Raksha Kumaraswamy, Anurag Wazalwar, Tushar Khot, Jude W. Shavlik, and Sriraam Natarajan. Anomaly detection in text: The value of domain knowledge. In Ingrid Russell and William Eberle, editors, *FLAIRS Conference*, pages 225–228. AAAI Press, 2015.

[21] Marcin Malec, Tushar Khot, James Nagy, Erik Blasch, and Sriraam Natarajan. Inductive logic programming meets relational databases: An application to statistical relational learning. In *ILP*, 2016.

[22] Emily Mallory, Ce Zhang, Christopher Ré, and Russ Altman. Large-scale extraction of gene interactions from full text literature using DeepDive. *Bioinformatics*, 2015.

[23] Emily K Mallory et al. Large-scale extraction of gene interactions from full text literature using deep-dive. *Bioinformatics*, 2015.

[24] Ndapandula Nakashole et al. Scalable knowledge harvesting with high precision and high recall. In *WSDM*, 2011.

[25] Sriraam Natarajan, Kristian Kersting, Tushar Khot, and Jude Shavlik. Boosting in the presence of missing data. *Boosted Statistical Relational Learners SpringerBriefs in Computer Science*, page 3948, 2014.

[26] Sriraam Natarajan, Tushar Khot, Kristian Kersting, Bernd Gutmann, and Jude Shavlik. Gradient-based boosting for statistical relational learning: The relational dependency network case. *Mach Learn Machine Learning*, 86(1):2556, Oct 2011.

[27] Sriraam Natarajan, Jose Picado, Tushar Khot, Kristian Kersting, Christopher Re, and Jude Shavlik. Effectively creating weakly labeled training examples via approximate domain knowledge. *Inductive Logic Programming Lecture Notes in Computer Science*, page 92107, 2013.

[28] Feng Niu et al. Elementary: Large-scale knowledge-base construction via machine learning and statistical inference. *Int. J. Semantic Web Inf. Syst.*, 2012.

[29] Feng Niu, Ce Zhang, Christopher Ré, and Jude W. Shavlik. DeepDive: Web-scale knowledge-base construction using statistical learning and inference. In *VLDS*, 2012.

[30] Feng Niu, Ce Zhang, Christopher Ré, and Jude W. Shavlik. Elementary: Large-scale knowledge-base construction via machine learning and statistical inference. *Int. J. Semantic Web Inf. Syst.*, 2012.

[31] Feng Niu, Ce Zhang, Christopher Ré, and Jude W. Shavlik. Scaling inference for Markov logic via dual decomposition. In *ICDM*, 2012.

[32] Phillip Odom, Vishal Bangera, Tushar Khot, David Page, and Sriraam Natarajan. Extracting adverse drug events from text using human advice. *Artificial Intelligence in Medicine Lecture Notes in Computer Science*, page 195204, 2015.

[33] Phillip Odom, Tushar Khot, Reid Porter, and Sriraam Natarajan. Knowledge-based probabilistic logic learning. In *AAAI*, pages 3564–3570, 2015.

[34] Shanan Peters, Ce Zhang, Miron Livny, and Christopher Ré. A machine-compiled macroevolutionary history of Phanerozoic life. *PLoS One*, 2014.

[35] Shanan E Peters et al. A machine reading system for assembling synthetic Paleontological databases. PloS ONE, 2014.

[36] Hoifung Poon and Pedro Domingos. Joint inference in information extraction. In *AAAI*, 2007.

[37] Warren Shen et al. Declarative information extraction using datalog with embedded extraction predicates. In *VLDB*, 2007.

[38] Jaeho Shin, Sen Wu, Feiran Wang, Christopher De Sa, Ce Zhang, and Christopher Ré. Incremental knowledge base construction using DeepDive. *PVLDB*, 2015. **Invited to VLDB Journal "Best of VLDB 2015"**.

[39] A. Soni, D. Viswanathan, N. Pachaiyappan, and S. Natarajan. A comparison of weak supervision methods for knowledge base construction. In *Automated Knowledge Base Construction (AKBC) Workshop at NAACL*, 2016.

[40] Srikrishna Sridhar, Stephen J. Wright, Christopher Ré, Ji Liu, Victor Bittorf, and Ce Zhang. An approximate, efficient LP solver for LP rounding. In *NIPS*, 2013.

[41] Fabian M Suchanek et al. SOFIE: A self-organizing framework for information extraction. In *WWW*, 2009.

[42] Shuo Yang, Tushar Khot, Kristian Kersting, Gautam Kunapuli, Kris Hauser, and Sriraam Natarajan. Learning from imbalanced data in relational domains: A soft margin approach. *2014 IEEE International Conference on Data Mining*, 2014.

[43] Alexander Yates et al. TextRunner: Open information extraction on the Web. In *NAACL*, 2007.

[44] Ce Zhang et al. GeoDeepDive: statistical inference using familiar data-processing languages. In *SIGMOD*, 2013.

[45] Ce Zhang, Vidhya Govindaraju, Jackson Borchardt, Tim Foltz, Christopher Ré, and Shanan Peters. GeoDeepDive: statistical inference using familiar data-processing languages. In *SIGMOD*, 2013.

[46] Ce Zhang, Arun Kumar, and Christopher Ré. Materialization optimizations for feature selection workloads. In *SIGMOD*, 2014. **SIGMOD 2014 Best Paper Award**.

[47] Ce Zhang, Feng Niu, Christopher Ré, and Jude W. Shavlik. Big data versus the crowd: Looking for relationships in all the right places. In *ACL*, 2012.

[48] Ce Zhang and Christopher Ré. Towards high-throughput Gibbs sampling at scale: a study across storage managers. In *SIGMOD*, 2013.

[49] Ce Zhang and Christopher Ré. Dimmwitted: A study of main-memory statistical analytics. *PVLDB*, 2014.

[50] Yingbo Zhou, Utkarsh Porwal, Ce Zhang, Hung Q. Ngo, Long Nguyen, Christopher Ré, and Venu Govindaraju. Parallel feature selection inspired by group testing. In *NIPS*, 2014.

[51] Jun Zhu et al. StatSnowball: A statistical approach to extracting entity relationships. In *WWW*, 2009.

[52] Yuke Zhu, Ce Zhang, Christopher Ré, and Li Fei-Fei. Building a large-scale multimodal knowledge base for visual question answering. *ArXiv e-prints*, 2015.

A Publications Resulting from this Cooperative Agreement.

A.1 Books

1. Sriraam Natarajan, Tushar Khot, Kristian Kersting and Jude Shavlik. Boosted Statistical Relational Learners: From Benchmarks to Data-Driven Medicine. *SpringerBriefs in Computer Science, ISBN: 978-3-319-13643-1,* 2015.

A.2 Dissertations

1. Tushar Khot. *Efficient Learning of Statistical Relational Models*. PhD thesis, Department of Computer Sciences, University of Wisconsin-Madison, 2014.

2. Ce Zhang. *DeepDive: A Data Management System for Automatic Knowledge Base Construction*. PhD thesis, Department of Computer Sciences, University of Wisconsin-Madison, 2015.

A.3 Conference Papers

1. Sriraam Natarajan, Philip Odom, Saket Joshi, Tushar Khot, Kristian Kersting and Prasad Tadepalli. Accelerating Imitation Learning in Relational Domains via Transfer by Initialization. In *Proceedings of the International Conference on Inductive Logic Programming (ILP)*, 2013.

2. Baidya Saha, Gautam Kunapuli, Nilanjan Ray, Joseph Maldijan and Sriraam Natarajan. AR-Boost: Reducing Overfitting by a Robust Data-Driven Regularization Strategy. In *Proceedings of the European Conference on Machine Learning, (ECMLPKDD),* 2013.

3. Shuo Yang and Sriraam Natarajan. Knowledge Intensive Learning: Combining Qualitative Constraints with Causal Independence for Parameter Learning in Probabilistic Models. In *Proceedings of the European Conference on Machine Learning, (ECMLPKDD)*, 2013.

4. Gautam Kunapuli, Philip Odom, Jude Shavlik and Sriraam Natarajan. Guiding Autonomous Agents to Better Behaviors through Human Advice. In *Proceedings of the IEEE International Conference on Data Mining (ICDM),* 2013.

5. Ce Zhang and Christopher Ré. Towards high-throughput Gibbs sampling at scale: a study across storage managers. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2013, New York, NY, USA, June 22-27, 2013,* pages 397–408, 2013.

6. Vidhya Govindaraju, Ce Zhang, and Christopher Ré. Understanding tables in context using standard NLP toolkits. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 2: Short Papers, pages 658–664, 2013.

7. Sriraam Natarajan, Jose Manuel Picado Leiva, Tushar Khot, Kristian Kersting, Christopher Re and Jude Shavlik. Effectively creating weakly labeled training examples via approximate domain knowledge. In *Proceedings of the International Conference on Inductive Logic Programming, (ILP),* 2014.

8. Tushar Khot, Sriraam Natarajan and Jude Shavlik. Relational One-Class Classification: A Non-Parametric Approach. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence (AAAI),* 2014.

9. Shuo Yang, Tushar Khot, Kristian Kersting, Gautam Kunapuli, Kris Hauser and Sriraam Natarajan. Learning from Imbalanced Data in Relational Domains: A Soft Margin Approach. In *Proceedings of the International Conference on Data Mining (ICDM),* 2014.

10. Ce Zhang and Christopher Re. Dimmwitted: A study of main-memory statistical analytics. Proceedings of Very Large Database Endowment *PVLDB*, 7(12):1283–1294, 2014.

11. Phillip Odom, Tushar Khot, Reid Porter, and Sriraam Natarajan. Knowledge-Based Probabilistic Logic Learning. In *Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI),* 2015.

12. Phillip Odom, Vishal Bangera, Tushar Khot, David Page and Sriraam Natarajan. Extracting Adverse Drug Events from Text using Human Advice. In *Proceedings of the Artificial Intelligence in Medicine (AIME),* 2015.

13. Raksha Kumaraswamy, Phillip Odom, Kristian Kersting, David Leake, and Sriraam Natarajan. Transfer Learning via Relational Type Matching. In *Proceedings of the International Conference on Data Mining (ICDM)*, 2015.

14. Jaeho Shin, Sen Wu, Feiran Wang, Christopher De Sa, Ce Zhang, and Christopher Ré. Incremental knowledge base construction using deepdive. Proceedings of Very Large Database Endowment *PVLDB*, 8(11):1310–1321, 2015 (**Invited to VLDB Journal ”Best of VLDB 2015”**).

15. Jaeho Shin, Christopher Ré, and Michael J. Cafarella. Mindtagger: A demonstration of data labeling in knowledge base construction. Proceedings of Very Large Database Endowment *PVLDB*, 8(12):1920– 1923, 2015.

16. Christopher De Sa, Ce Zhang, Kunle Olukotun, and Christopher Ré. Taming the wild: A unified analysis of hogwild-style algorithms. In *Advances in Neural Information Processing Systems (NIPS) December 7-12, 2015, Montreal, Quebec, Canada*, pages 2674–2682, 2015.

17. Christopher De Sa, Ce Zhang, Kunle Olukotun, and Christopher Ré. Rapidly mixing gibbs sampling for a class of factor graphs using hierarchy width. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 3097–3105, 2015.

18. Mahmoud Abo Khamis, Hung Q. Ngo, Christopher Ré, and Atri Rudra. Joins via geometric resolutions: Worst-case and beyond. In Proceedings of the 34th ACM Symposium on Principles of Database Systems, (PODS) 2015, Melbourne, Victoria, Australia, May 31 - June 4, 2015, pages 213–228, 2015.

19. Mayukh Das, Yuqing Wu, Tushar Khot, Kristian Kersting and Sriraam Natarajan. Scaling Lifted Probabilistic Inference and Learning Via Graph Databases. In *Proceedings of the SIAM International Conference on Data Mining (SDM)*, 2016.

20. Manas Joglekar and Christopher Ré. It's all a matter of degree: Using degree information to optimize multiway joins. In *19th International Conference on Database Theory, (ICDT) 2016, Bordeaux, France, March 15-18, 2016*, pages 11:1–11:17, 2016.

21. Marcin Malec, Tushar Khot, James Nagy, Erik Blasch and Sriraam Natarajan. Inductive Logic Programming meets Relational Databases: An Application to Statistical Relational Learning. In *Proceedings of the Inductive Logic Programming (ILP)*, 2016 (**Best paper award**).

22. Ameet Soni, Dileep Viswanathan, Jude Shavlik and Sriraam Natarajan. Learning Relational Depen- dency Networks for Relation Extraction. Under Review, 2016.

23. Navdeep Kaur, Gautam Kunapuli, Tushar Khot, Kristian Kersting, William Cohen and Sriraam Natara- jan. Relational Restricted Boltzmann Machines: A Probabilistic Logic Learning Approach. In *Pro- ceedings of the Inductive Logic Programming (ILP)*, 2017.

A.4 Journal Papers

1. Christopher Ré, Amir Abbas Sadeghian, Zifei Shan, Jaeho Shin, Feiran Wang, Sen Wu, and Ce Zhang. Feature engineering for knowledge base construction. *IEEE Data Eng. Bull.*, 37(3):26–40, 2014.

2. Shanan Peters, Ce Zhang, Miron Livny, and Christopher Ré. A machine-compiled macroevolutionary history of phanerozoic life. *PLoS One*, 2014.

3. Tushar Khot, Sriraam Natarajan, Kristian Kersting, Bernd Gutmann and Jude Shavlik. Gradient- based Boosting for Statistical Relational Learning: The Markov Logic Network and Missing Data Cases. *Machine Learning Journal*, 2015.

4. Emily K. Mallory, Ce Zhang, Christopher Ré, and Russ B. Altman. Large-scale extraction of gene interactions from full-text literature using deepdive. *Bioinformatics*, 32(1):106–113, 2016.

5. Ce Zhang, Arun Kumar, and Christopher Ré. Materialization optimizations for feature selection workloads. *ACM Trans. Database Syst.*, 41(1):2, 2016.

A.5 Book Chapter

1. Sriraam Natarajan, Ameet Soni, Anurag Wazalwar, Dileep Viswanathan and Kristian Kersting. Deep Distant Supervision: Learning Statistical Relational Models for Weak Supervision in Natural Lan- guage Extraction. In *Proceedings of the Morik Festschrift, LNAI 9580*, 2016

A.6 Workshop Papers

1. Sriraam Natarajan, Jose Picado, Tushar Khot, Kristian Kersting, Christopher Re and Jude Shavlik. Shavlik.Using Commonsense Knowledge to Automatically Create (Noisy) Training Examples from Text. In *Proceedings of the International Workshop on Statistical Relational AI*, 2012.

2. Mayukh Das, Yuqing Wu, Tushar Khot, Kristian Kersting and Sriraam Natarajan. Graph-based Ap- proximate Counting for Relational Probabilistic Models. In *Proceedings of theInternational Workshop on Statistical Relational AI (StarAI)*, 2015.

3. Benny Kimelfeld and Christopher Re. A database framework for classifier engineering. In Pro- ceedings of the 9th Alberto Mendelzon International Workshop on Foundations of Data Management, Lima, Peru, May 6 - 8, 2015., 2015.

4. Stefan Hadjis, Firas Abuzaid, Ce Zhang, and Christopher Ré. Caffe con troll: Shallow ideas to speed up deep learning. In *Proceedings of the Fourth Workshop on Data analytics in the Cloud, DanaC 2015, Melbourne, VIC, Australia, May 31 - June 4, 2015*, pages 2:1–2:4, 2015.

5. A. Soni, D. Viswanathan, N. Pachaiyappan, and S. Natarajan. A comparison of weak supervision methods for knowledge base construction. In *Automated Knowledge Base Construction (AKBC) Workshop at NAACL*, 2016.

6. Dileep Viswanathan, Ameet Soni, Jude Shavlik and Sriraam Natarajan. Learning Relational Depen- dency Networks for Relation Extraction, *International Workshop on Statistical Relational AI (StarAI)* 2016.

7 LIST OF ACRONYMS

AAAI – Association for the Advancement of Artificial Intelligence
ADE – Adverse Drug Events
AFRL – Air Force Research Laboratory
AUC – Area Under the Curve
DARPA – Defense Advanced Research Projects Agency
DEFT – Deep Exploration and Filtering of Text
DN – Dependency Networks
DPR – Root of dependency Path tree
EM – Expectation–Maximization
FGB – Functional-Gradient Boosting
FOL – First Order Logic
ICDM – International Conference Data Mining
IJCAI – International Joint Conference on Artificial Intelligence
ILP – International Conference on Inductive Logic Programming
KB – Knowledge Base
KBP – Knowledge Base Population
KBC – Knowledge Base Construction
LTL - Language-biased Transfer Learning
MAP – maximum a posteriori
MLN – Markov Logic Network
M^{2}T – Mode-Matching Trees
NE – Named Entity
NELL – Never Ending Language Learning
NIST – National Institute of Standards and Technology
NLP – Natural Language Processing
NUMC - non-uniform memory access
OMOP - Observational Medical Outcomes Partnership
POS – Part of Speech
RBM - Restricted Boltzman Machines
RDN – Relational Dependency Network
RF – Relation Factory
RFGB – Relational Functional-Gradient Boosting
ROC – Receiver operating characteristic
RRT – Relational Regression Trees
SDM – SIAM International conference on Data Mining
SRL – Statistical Relational Learning
SVM – Support Vector Machine
TAC – Text Analysis Conference
USAF – United States Air Force